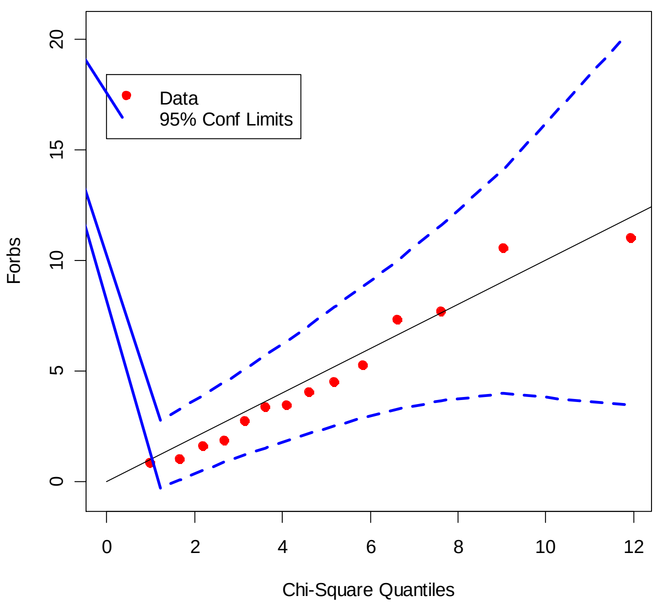
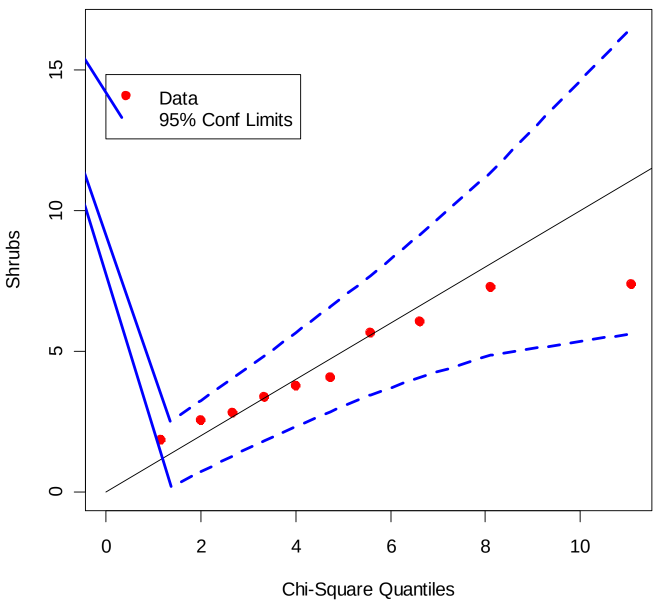
Example: Land Cover. *Land cover is classified into two functional groups: forbs (1) and shrubs (2). The data was collected at two sites: Colorado and Wisconsin. For now, we'll pool this data together (data described in Quinn & Keough, Experimental Design and Analysis for Biologists). Online as reich.csv (the variable 'FUNCTION' indicates forb(1) vs shrub (2) – we'll deal with other variables later). The following variables were also measured:*

- *N - concentration of nitrogen*
- *AMASS - - mass-based net photosynthetic capacity*
- *AAREA- area-based net photosynthetic capacity*
- *GS - leaf diffuse conductance at photosynthetic capacity*
- *LSLA - log10 transformation of specific leaf area*

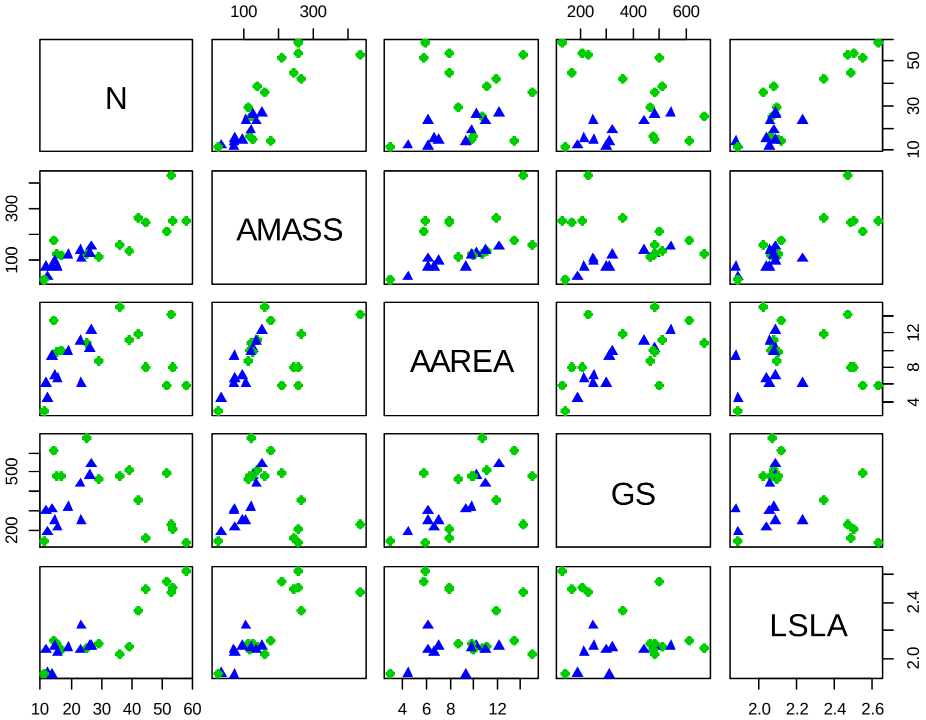
We use discriminant analysis to discriminate forbs and shrubs based on the five other measured variables

Discriminant Analysis in R. First, load the MASS package. Then use the `lda()` function. See example up on CANVAS home page.


Chi-Square Quantiles for Forbs


Chi-Square Quantiles for Shrubs

Assumption of multivariate normality within each group seems reasonable. Next we look at where the groups lie in two dimensions at a time:



*Seems likely that the covariances matrices are **NOT** the same between groups. We’ll ignore this for the moment and talk about how to fix this later. As a confirmation of this problem, we look at the sample standard deviations in each group: doesn’t look equal, especially for Nitrogen, AMASS, LSLA*

```
                N  AMASS  AAREA     GS    LSLA
 Forbes  16.41  99.48   3.44  180.71   0.24
Shrubs    5.78  36.40   2.56  118.72   0.10
```

Calculate Box’s M statistic (requires `biotools` *package) – we reject the null hypothesis (which is not what we were hoping for . . .)*

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  forbs[, 4:8]
Chi-Sq (approx.) = 42.135, df = 15, p-value = 0.0002142
```

At this point, we're basically doing Multivariate Analysis of Variance to get these results (which is the next topic) –

```
Response N:
Df Sum Sq Mean Sq F value Pr(>F)
forbs[, 2] 1 1551.2 1551.2 8.9806 0.006644 **

Response AMASS:
Df Sum Sq Mean Sq F value Pr(>F)
forbs[, 2] 1 46997 46997 7.3551 0.01273 *

Response AAREA:
Df Sum Sq Mean Sq F value Pr(>F)
forbs[, 2] 1 11.340 11.340 1.1751 0.2901

Response GS:
Df Sum Sq Mean Sq F value Pr(>F)
forbs[, 2] 1 20090 20090 0.8016 0.3803
```

```
Response LSLA:
Df Sum Sq Mean Sq F value Pr(>F)
forbs[, 2] 1 0.22443 0.22443 5.8932 0.02384
```

Univariate Analyses show difference for Nitrogen, AMASS, and LSLA.

```
Df Wilks approx F num Df den Df Pr(>F)
forbs[, 2] 1 0.58777 2.52487 5 18 0.06707
```

Questionable significance at the multivariate level!

We’ll return to this example in a bit when we have a few more tools . . .

__Example: Depression__. One in seven Americans will experience clinical depression at some point during their lifetime. A study of 297 people was conducted at the UCLA Social Science Research institute (data described in Computer-Aided Multivariate Analysis, Clark and May, 2004). Respondents were categorized using the CESD Depression Index of the NIMH. Values > 16 are considered to indicate clinical depression (scale is from 0 to 60). The variable `CASES` *(1=Clinical Depression, 0=No Clinical Depression). Data is online as* `DEPRESSION.CSV`. *The following variables were also measured:*

- *Education (1-7 scale with 7 being the most)*
- *Income (thousands dollars per year)*
- *Health (1-4 scale from Excellent=1 to Poor=4)*
- *Age (in years)*

We use discriminant analysis to discriminate between individuals with clinical depression and those without. R script is up on CANVAS.

We'll discuss results in class . . .

Classification

Classification is considered with Discriminant Analysis for two reasons:

1. Want to evaluate how well our discrimination function worked on the **data** where **group membership is known**
2. Want to evaluate how well our discriminant function would work on **new data** where **group membership is unknown.**

Classification can be a separate topic from Discriminant Analysis, or can be part of DA. We consider

- Cut-off method
- Decision Theory
- Classification functions

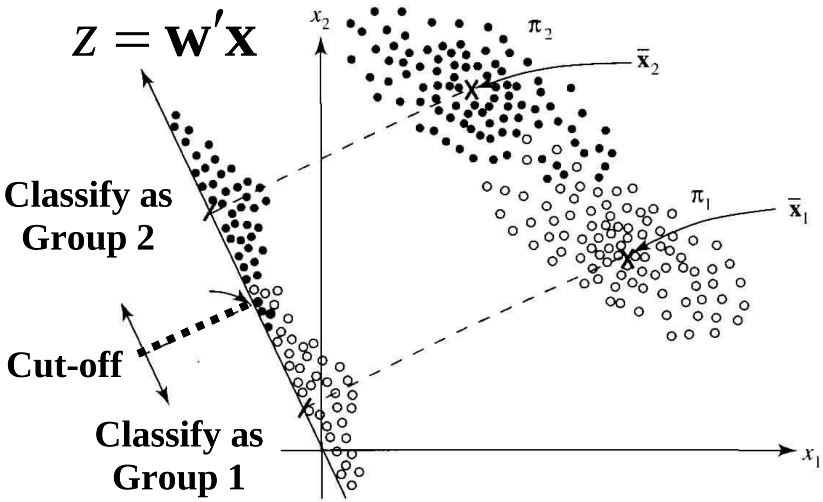
Cut-off method

The discriminant function:

$$z = \mathbf{w}'\mathbf{x} = w_1 x_1 + w_2 x_2 + \ldots + w_p x_p$$

- This function reduces data to a **univariate** problem.
- Choose a **cut-off** value along this new axis to divide into two regions.

- Calculate **discriminant scores (these are the z values!!!)**: for each observation, plug in values for $x_1, x_2 ..., x_p$. **The scores are the projection of the points onto the direction of maximum discrimination!**



$$z = \mathbf{w}'\mathbf{x}$$

Classify as Group 2

Cut-off

Classify as Group 1

- Observations with scores above the cut-off value are classified in one group, all others are classified in the other group.
- Cut-off value usually chosen to **minimize the number of misclassifications in the sample data.**

- Usual Cut-off value for **two groups** of **equal samplesize** (assumes equal covariances matrices between groups):

$$\text{cut}-\text{off value} = \frac{\bar{z}_1 + \bar{z}_2}{2}$$

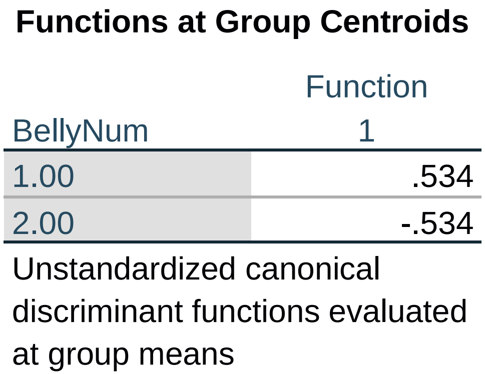
where $\bar{z}_g$ is the mean discriminant score in group g

- Equivalent value for unequal sample sizes

$$\text{cut}-\text{off value} = \frac{n_1\bar{z}_1 + n_2\bar{z}_2}{n_1 + n_2}$$

where n_g is the sample size of group g

Example: Sneetches (in SPSS). Equal sized groups** (45 star and 45 plain), cut-off value based on unstandardized data (with added constant) is **zero!

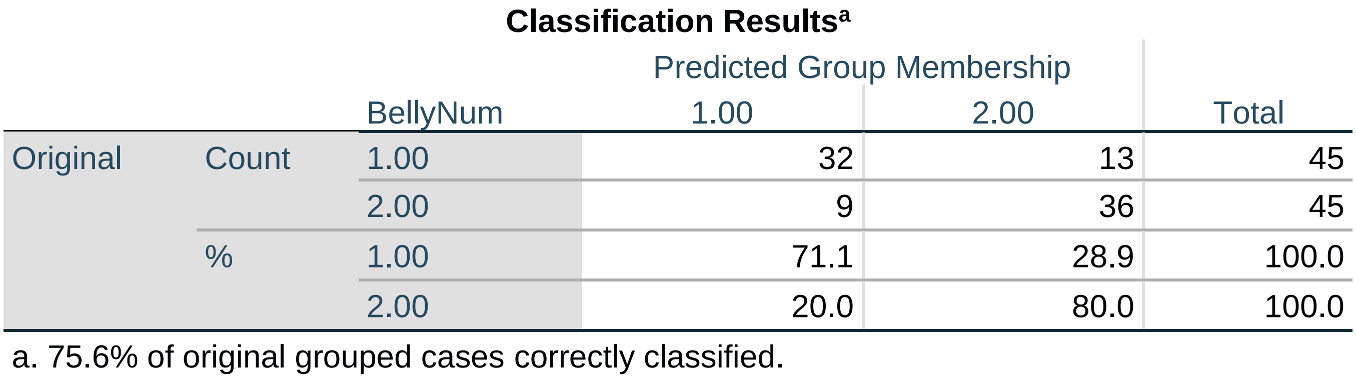


Functions at Group Centroids

BellyNum	Function 1
1.00	.534
2.00	-.534

Unstandardized canonical discriminant functions evaluated at group means

SPSS also provides a classification matrix:
a summary of how well classification worked: called a confusion matrix *(click on* `Classify` *and then* `Summary Table`*)*



Classification Results[a]

			Predicted Group Membership		
		BellyNum	1.00	2.00	Total
Original	Count	1.00	32	13	45
		2.00	9	36	45
	%	1.00	71.1	28.9	100.0
		2.00	20.0	80.0	100.0

a. 75.6% of original grouped cases correctly classified.

Example: Land Cover. *Classification Results:*

Classification Results[a]

			Predicted Group Membership		
		FUNCTION	1	2	Total
Original	Count	1	11	3	14
		2	2	8	10
	%	1	78.6	21.4	100.0
		2	20.0	80.0	100.0

a. 79.2% of original grouped cases correctly classified.

Example: Depression. Unequal Sample Sizes *(244 not-depressed, 50 depressed). Slightly better than guessing!*

Classification Results[a]

			Predicted Group Membership		
		CASES	0	1	Total
Original	Count	0	157	87	244
		1	15	35	50
	%	0	64.3	35.7	100.0
		1	30.0	70.0	100.0

a. 65.3% of original grouped cases correctly classified.

Now: what if

Example: *in the depression example, only about 1 in 6 people was actually clinically depressed. It might be reasonable to prefer classification as 'not depressed' in the absence of other information:* **Use prior information in classification decisions**

Example: *We have to decide whether to take action to prevent invasion by snakefish. Which is worse – to take preventative measures when there was no danger of invasion, or to not take action and have snakefish invade?*

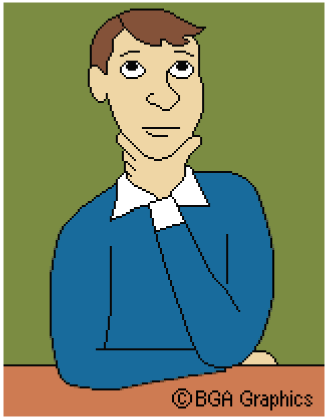
Which is worse – to replace a critical part on an airplane that was fine or to fail to replace it when it is faulty?

Account for costs of different misclassifications

***Example**: What if our data doesn’t look anything like a multivariate normal distribution?*

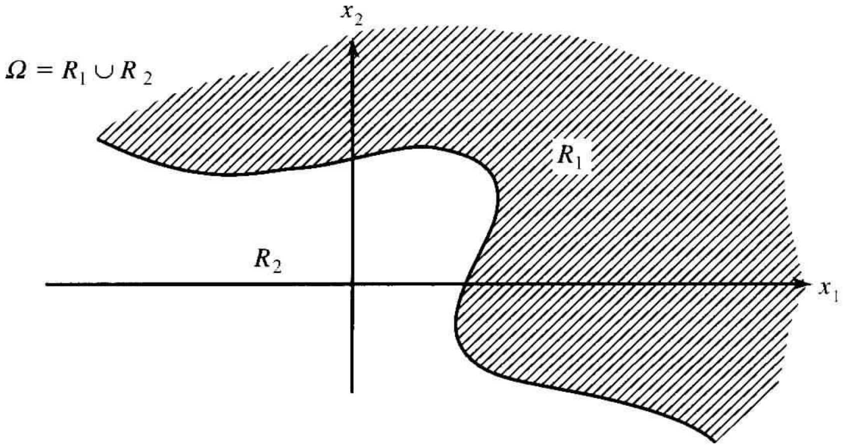
These questions lead to broader view of classification based on . . .

Decision Theory



- Let Ω (Omega) be the p-dimensional sample space of all possible multivariate observations $\mathbf{x}$
- Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the **probability density functions** associated with each group defined over Ω (think two overlapping mountain ranges).
- Partition Ω into two regions R_1 and R_2 such that $R_2 = \Omega - R_1$ (i.e. every point in Ω is either in R_1 or in R_2)
- For observations $\mathbf{x}$ in R_1, classify an observation as in group 1, otherwise, classify in group 2

An example for p=2 dimensions:

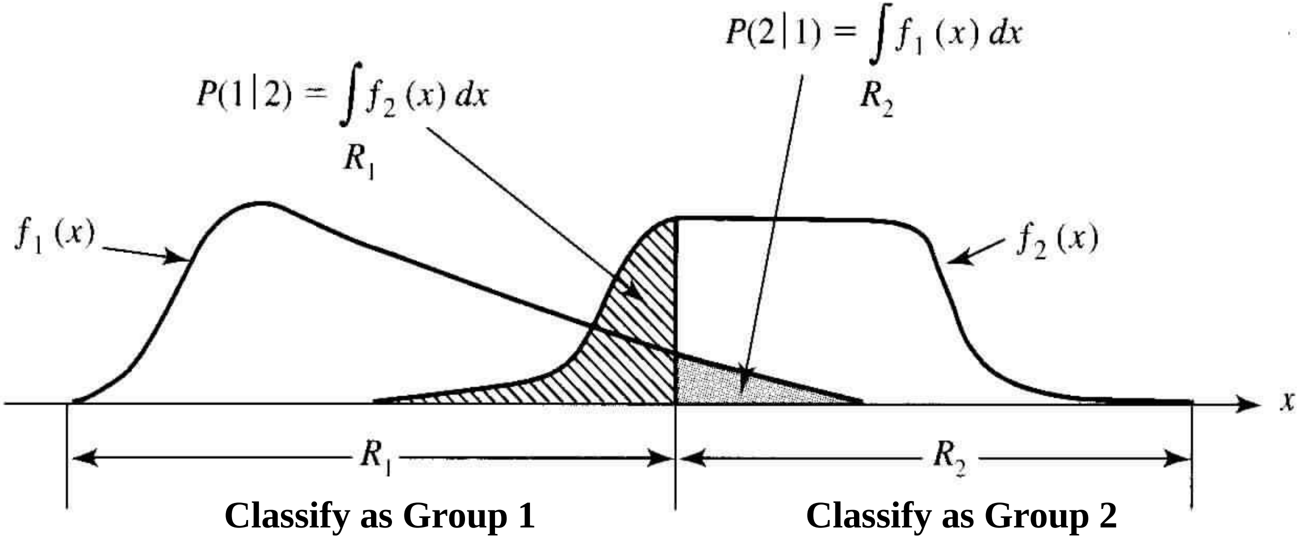


Misclassification probabilities: $P(2|1)$ = conditional probability of misclassifying an object as belonging to group 2 when it is really from group 1. Use Calculus - 'area under the curve' in p -dimensions:

$$P(2|1) = \int_{R_2} f_1(\mathbf{x})\,d\mathbf{x}$$

Similarly, get $$P(1|2) = \int_{R_1} f_2(\mathbf{x})\,d\mathbf{x}$$

An example in p=1 dimension:



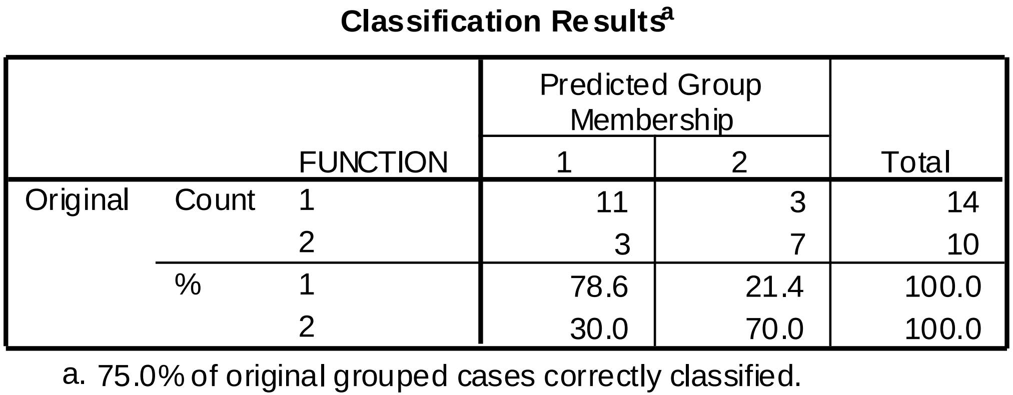
Prior probabilities

- Before performing classification, may believe that relative proportion of each group is not equal
- Most common case: a simple random sample yields unequal sample sizes: indicates that rates of occurrence of two groups are not identical
- Theory may also indicate unequal group sizes (i.e. recessive genes, etc).
- Let p_1 and p_2 be prior probability of being in groups one and two such that $p_1 + p_2 = 1$

- **Overall probabilities of misclassification**:
  - ○ Probability observation is misclassified as in group 2 = $P(2|1)\,p_1$
  - ○ Probability observation is misclassified as in group 1 = $P(1|2)\,p_2$

Computer Notes: Specifying prior probabilities. In MINITAB, click on OPTIONS. In SPSS, click on CLASSIFY (can only base on group sizes). In SAS, use the PRIORS statement in PROC DISCRIM. In R use the `priors` option in the `lda()` function (in package MASS).

Example: Land Cover. *There are 14 forb plots and 10 shrub plots in our data, so maybe should prefer forbs by default. Here is classification with priors equal to observed sample sizes: classification gets slightly worse! (one more shrub is misclassified)*



Classification Results[a]

			Predicted Group Membership		
		FUNCTION	1	2	Total
Original	Count	1	11	3	14
		2	3	7	10
	%	1	78.6	21.4	100.0
		2	30.0	70.0	100.0

a. 75.0% of original grouped cases correctly classified.

Misclassification Costs

- Different misclassifications may have unequal costs (SpaceX rocket launch, tumor types (malignant/benign)).
- May want classification rule to account for misclassification costs
- Matrix of misclassifications costs:

		Classify as:	
		Group 1	Group 2
True	Group 1	0	$c(2\|1)$
Population:	Group 2	$c(1\|2)$	0

- **Expected Cost of misclassification (ECM):**

$$ECM = P(2|1)\,p_1 c(2|1) + P(1|2)\,p_2 c(1|2)$$

Goal: pick regions *R₁* and *R₂* to minimize ECM.

Fact (a bit of algebra): To minimize ECM, choose regions

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

This is simpler for equal priors and equal misclassification costs:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

i.e., R_1 is the region where the density function for group 1 is larger than the density function for group 2, and vice-versa.

To summarize: classification regions can be chosen based on any distributions and may incorporate misclassification costs and prior distribution information.

Computer Notes: Including misclassification costs. For all programs, just incorporate costs into prior distributions: (hard to do this in SPSS . .)

$$ECM = P(2|1)\underbrace{p_1 c(2|1)} + P(1|2)\underbrace{p_2 c(1|2)}$$

Make these the priors (say p_1^* *and* p_2^* *)*

Special Case: Two multivariate normal distributions with equal covariance matrices *(see Johnson and Wichern, p. 591-592, also Zelterman section 10.3)*

- Probability Density Function for multivariate normal:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}}$$

- Use formulas to minimize Expected Cost of Misclassification (ECM)

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \qquad R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_1 : \frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}{2}}}{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}{2}}} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Take logs:

$$R_1 : -\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

- Estimate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{2,} \boldsymbol{\Sigma}$ (unknown, true means and covariance) with $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_{2,} \mathbf{S}_{pooled}$ (the estimate of the common covariance matrix based on <u>both</u> samples)

- Algebra shows R_1 is the region of points $\mathbf{x}^*$ such that

$$\boxed{R_1 : \underset{1\times p}{(\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_2)'} \underset{p\times p}{\mathbf{S}^{-1}_{pooled}} \underset{p\times 1}{\mathbf{x}^*} - \frac{1}{2} \underset{1\times p}{(\bar{\mathbf{x}}_{\mathbf{1}} - \bar{\mathbf{x}}_2)'} \underset{p\times p}{\mathbf{S}^{-1}_{pooled}} \underset{p\times 1}{(\bar{\mathbf{x}}_{\mathbf{1}} + \bar{\mathbf{x}}_2)}}$$

This is just a scalar – a number!

$$\geq \boxed{\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]}$$

This is the CUTOFF VALUE!

and R_2 is all other points.

This is the estimated ECM rule for two normal populations with equal covariance matrices

Notes:

- The multivariate problem has been reduced to a univariate problem: This is a **LINEAR** function of $\mathbf{x}^*$
- This is an **estimated** minimum ECM rule – we don't know the true means and covariance matrix, we're using estimates.

NOW: If

- prior probabilities and classification costs are equal
- we define $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{pooled}^{-1} \mathbf{x} = \mathbf{w}'\mathbf{x} = z$ *(think rotation!)* which implies that

$$\frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{\bar{z}_1 + \bar{z}_2}{2}$$

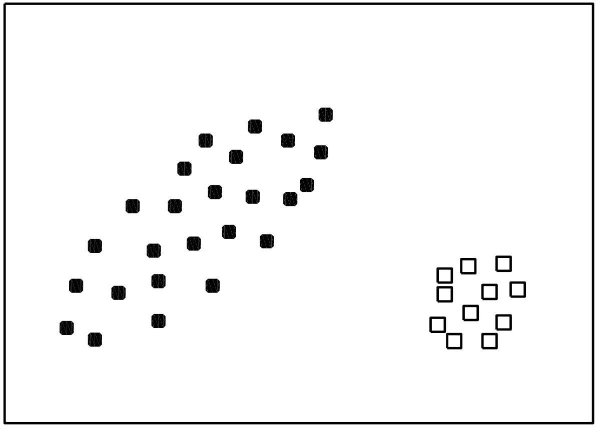
I.E.: Classification rule is R_1 is the region where

$$z \geq \frac{\bar{z}_1 + \bar{z}_2}{2}$$

Just the cut-off rule!

The Cut-Off Rule is equivalent to the ECM rule for two multivariate normal distributions with equal covariance matrices, equal prior distributions, and equal costs of misclassification.

Special Case : Two multivariate normal distributions with unequal covariance matrices



- Use formulas to minimize Expected Cost of Misclassification (ECM), and estimate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{2,} \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ (unknown, true means and covariances) with $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1, \mathbf{S}_2$

- Taking ratios, taking logs *(see Johnson and Wichern, p. 596-598)*, algebra then shows R_1 is the region of points $\mathbf{x}^*$ such that

$$R_1 : -\frac{1}{2}\mathbf{x}^{*\prime}\left(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1}\right)\mathbf{x}^* + \left(\bar{\mathbf{x}}_\mathbf{1}{}'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2{}'\mathbf{S}_2^{-1}\right)\mathbf{x}^* - k$$

$$\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

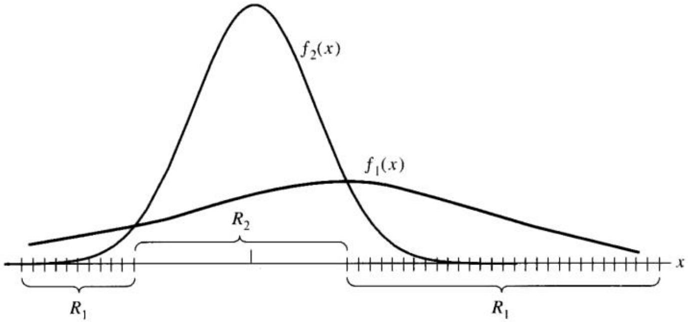
and R_2 is all other points. Here, k is a constant based on $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1, \mathbf{S}_2$

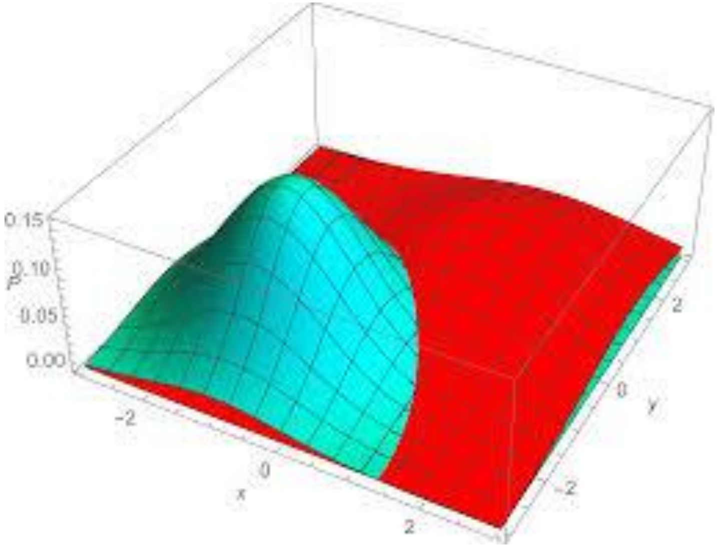
This is the estimated ECM rule for two normal populations with unequal covariance matrices

- This rule is a **quadratic** function of the original variables $\mathbf{x}$, not a linear function of $\mathbf{x}$. This is known as a **quadratic discriminant function**. Up to here, we've considered **linear** discriminant functions.

Notes on quadratic discrimination functions

- Used when data is multivariate normal but **covariance matrices are different between groups**
- Rejection region has an unusual shape: R_1 is disjoint
- **Not** to be used for non-normal data
- For cases of approximately equal covariance matrices, better to use linear discrimination because quadratic discrimination uses (many) more degrees of freedom estimating a second covariance matrix.





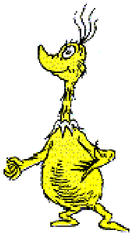
Computer Notes: Quadratic Discrimination. In MINITAB, click on Quadratic Discriminant Function. In SPSS, under Classify, click on Separate Covariance Matrices. In SAS, use the POOL=NO option. In R, use the `qda()` function in the MASS package (for **Q**uadratic **D**iscriminant **A**nalysis)

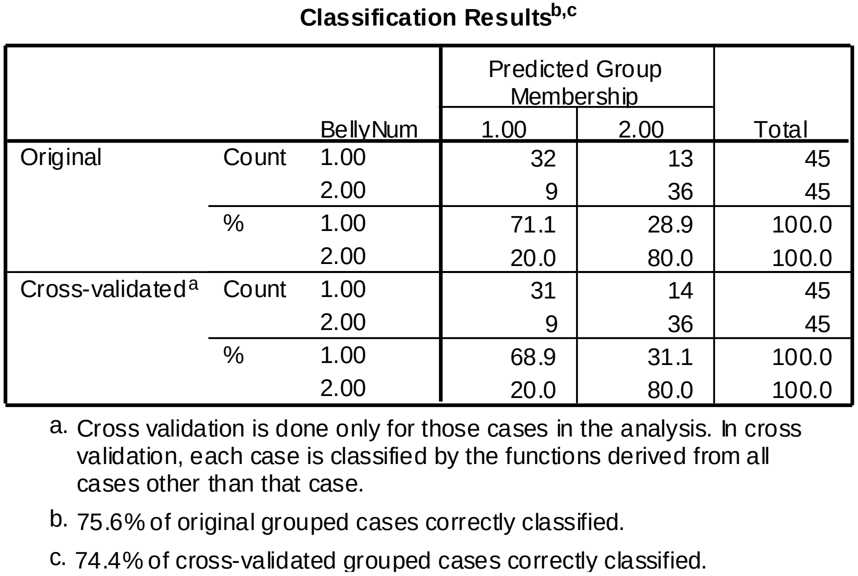
Leave-one-out Classification

- Known as Lachenbruch’s “holdout” procedure, jackknifing, cross-validation
- Use the following steps:
    1. In Group one, leave out one observation. Calculate classification function based on all remaining observations.
    2. Allocate the removed observation based on the decision rule created using all other data points
    3. Repeat 1) and 2) for all observations in both groups.

Computer Notes: Leave one out classification. Not available in MINITAB. In SPSS, under `Classify`, click on `Leave-one-out classification`. Not available with quadratic discrimination. In SAS, use the `CROSSVALIDATE` option. In R, use the `CV=TRUE` option in the `lda()` function. R also supports n-fold cross-validation (see code) – that is, leave out n-points at a time.

Example*:*
Classification results for Sneetches Data. Almost no change for Cross-classification data.





Classification Results[b,c]

			Predicted Group Membership		
		BellyNum	1.00	2.00	Total
Original	Count	1.00	32	13	45
		2.00	9	36	45
	%	1.00	71.1	28.9	100.0
		2.00	20.0	80.0	100.0
Cross-validated[a]	Count	1.00	31	14	45
		2.00	9	36	45
	%	1.00	68.9	31.1	100.0
		2.00	20.0	80.0	100.0

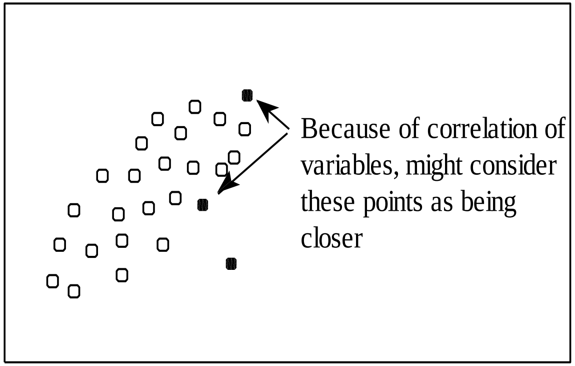
a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 75.6% of original grouped cases correctly classified.

c. 74.4% of cross-validated grouped cases correctly classified.

Mahalanobis Distance

- Remember that unlike Euclidean Distance, **Mahalanobis distance** accounts for correlation and differences of scale between variables.

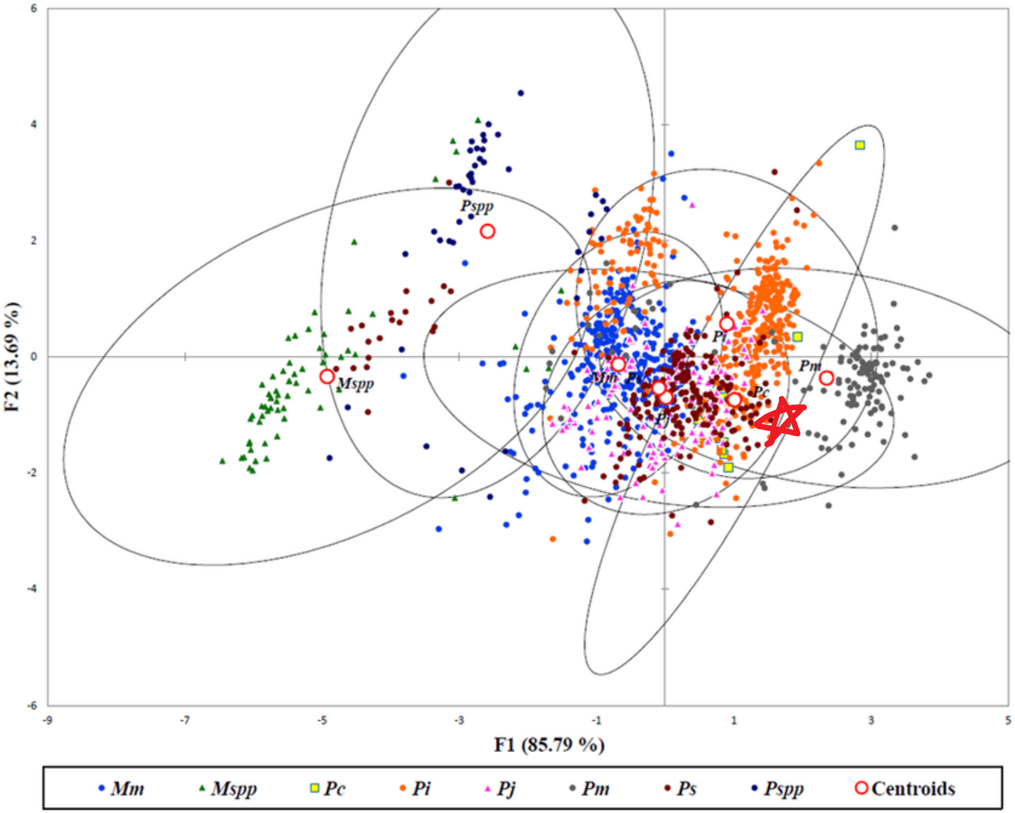


Because of correlation of variables, might consider these points as being closer

- Distance between points *i* and *j* is defined as $MD_{ij} = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_j\right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)}$
- When using the cutoff method, this is also equivalent to choosing the group whose centroid is the smallest Mahalanobis distance way.

Example (2019): *Discrimination of Penaeid shrimps in Malindi-Ungwana Bay based on MH distances.*

Think about which centroid is actually closest to the hypothetical shrimp at the red star – in particular, think about how far the point is from each 95% outer confidence boundary.



***Example**: Case by case classification of Sneetches data*

		Actual Group	Highest Group					Second Highest Group			Discriminant Scores
			Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
	Case #			p	df						
Original	1	1	2	.760	1	.561	.093	1	.439	.582	-.229
	2	1	1	.934	1	.659	.007	2	.341	1.325	.617
	3	1	2	.221	1	.867	1.499	1	.133	5.255	-1.758
	4	1	1	.967	1	.649	.002	2	.351	1.230	.575
	88	2	2	.876	1	.599	.024	1	.401	.831	-.378
	89	2	2	.326	1	.835	.966	1	.165	4.205	-1.517
	90	2	2	.722	1	.547	.127	1	.453	.507	-.178
Cross-validated	1	1	2	.926	2	.564	.155	1	.436	.672	
	2	1	1	.591	2	.652	1.052	2	.348	2.310	
	3	1	2	.320	2	.895	2.280	1	.105	6.565	
	4	1	1	.926	2	.647	.155	2	.353	1.363	
	88	2	2	.958	2	.598	.085	1	.402	.879	
	89	2	2	.528	2	.831	1.276	1	.169	4.465	
	90	2	2	.861	2	.545	.299	1	.455	.660	